Title: Audio Signal Processing Using Improved Perceptual Model

Field of the Invention

The present invention relates to audio signal processing systems and methods, including such systems and methods for spatial shaping of noise content of such audio signals. More particularly, the present invention relates to methods and systems for shaping noise associated with audio signals to permit hiding such noise in bands of lower sensitivity for human auditory perception. Still more particularly, the present invention relates to noise shaping to improve audio coding, including reduced bit-rate coding.

Background of the Invention

It has long been known that the human auditory response can be masked by audio-frequency noise or by other-than-desired audio frequency sound signals. See, B. Scharf, "Critical Bands," Chap. 5 in J. V. Tobias, *Foundations of Modern Auditory Theory*, Academic Press, New York, 1970. While *critical bands*, as noted by Scharf, relate to many analytical and empirical phenomena and techniques, a central features of critical band analysis relates to the characteristic of certain human auditory responses to be relatively constant over a range of frequencies. In the cited Tobias reference, at page 162, one possible table of 24 critical bands is presented, each having an identified upper and lower cutoff frequency corresponding to certain behavior of human cochlea. In some contexts, these or related bands are described in terms of a Bark scale. The totality of the bands covers the audio frequency spectrum up to 15.5 kHz. Critical band effects have been used to advantage in designing coders for audio signals. See, for example, M. R. Schroeder et al, "Optimizing Digital Speech Coders By Exploiting Masking Properties of the Human Ear," Journal of the Acoustical Society of America, Vol. 66, pp. 1647-1652, December, 1979 and U.S. Patent Re 36,714 issued May 23, 2000 to J.D. Johnston and K. Brandenburg.

In particular, noise shaping techniques have been widely employed in many speech, audio and image applications such as coding (compression) to take advantage of noise masking techniques in critical bands. See generally, N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings* of *the IEEE, vol.* 81, October 1993. Other areas in which noise shaping has proven useful include data hiding and watermarking, as described, for example, in G. C. Langelaar, I.

Setyawan, and R. L. Lagendijk, "Watermarking digital image and video data," *IEEE Signal Processing Magazine*, 2000.

One purpose of such prior techniques is to shape noise to be less perceptible (or not perceptible at all) in the final processed host signal. Many of these techniques shape noise by altering its spectrum, as, for example, using perceptual weighting filters in Code-Excited Linear Predictive (CELP) speech coders, and employing psychoacoustic models in audio coders. Some prior techniques developed for specific classes of applications have not proven useful over a wider range of applications.

Another approach known as *temporal noise shaping* (TNS) was described by J. Herre and J. D. Johnston in "Enhancing the performance of perceptual audio coding by using temporal noise shaping (TNS)," 101*st AES Convention, Los Angeles*, November 1996. The TNS method shapes the temporal structure of the quantization noise, instead of its spectrum as in many prior methods. One result of using the TNS approach is to effectively reduce the so-called *pre-echo* problem well known in audio coding that arises from the spread of quantization noise in the time domain within a transform window. In another aspect, TNS has proven useful in processing of certain signals having dominant pitch components. Importantly, TNS has greatly contributed to the high performance of MPEG Advanced Audio Coder (AAC). See, for example, J. D. Johnston, S. R. Quackenbush, G. A. Davidson, K. Brandenburg, and J. Herre, "MPEG audio coding," in *Wavelet, subband and block transforms in communications and multimedia* (A. N. Akansu and M. J. Medley, eds.), ch. 7, pp. 207-253, Kluwer Academic Publishers, 1999.

As noted above, prior noise shaping techniques have operated on signals in frequency bands corresponding roughly to respective frequency bands occurring in the human cochlea (*i.e.*, cochlea filter bands). Particular processing operations are typically based, at least in part, on an assumed model for human hearing. While many such models have proven useful in providing a basis for noise shaping purposes, nevertheless shortcomings have been discerned when applying various prior models.

Thus, for example, prior modeling of hearing has in some cases been based, at least in part, on processing based on the *tonal* and *noise-like* characteristics of input signals to determine a noise threshold, *i.e.*, a signal level below which noise will be masked. See, for example, U.S. Patent 5,341,457issued August 24, 1994 to J.L. Hall II

and J.D. Johnston. Often, it proves advantageous to characterize this noise-to-signal ration as a Noise Masking Ratio (NMR). However, as noted, *e.g.*, in U.S. Patent 5,699,479 issued December 16, 1997 to J.B. Allen, *et al.*, speech and music coders that exploit masking properties of an input sound to hide quantization noise are hampered by the difference in masking efficacy of tones and noise like signals when computing the masked threshold. In particular, developers of these coders seek to define the two classes of signals, as well as to identify the two classes in sub-bands of the input signal.

Summary of the Invention

Limitations of the prior art are overcome and a technical advance is made in accordance with the present invention described in illustrative embodiments herein.

In accordance with one illustrative embodiment based on psychoacoustic experiments, a perceptual model is introduced that is not based on evaluating the *noise-like vs. tonal* nature of the input signal. Rather, the masking ability of a signal in accordance with this illustrative embodiment is based on the (time domain) roughness of the envelope of an input signal in particular cochlea filter bands. In illustrative implementations, frequency domain techniques are used to develop necessary envelope and envelope roughness measures. A relationship is then advantageously developed between envelope roughness and NMR.

Thus, illustrative embodiments of the present invention provide systems and methods for realizing results of time domain masking techniques in the frequency domain, *i.e.*, for calculating NMRs for use in the frequency domain using time domain masking theory and improved processing techniques.

Illustrative coder embodiments of the present invention prove to be compatible with well-known AAC coding standards. Using present inventive techniques, standard MDCT coefficients can be efficiently quantized based on the present improved human perceptual model and improved processing techniques.

Brief Description of the Drawing

The above-summarized description of illustrative embodiments of the present invention will be more fully understood upon a consideration of the following detailed description and the attached drawing, wherein:

FIG. 1 is Bark scale plot of roughness of illustrative noise and pure tone input signals as determined in accordance with an aspect of the present invention.

FIG. 2 is a Bark scale plot of Noise Masking Ratio (NMR) for the illustrative noise and pure tone input signals reflected in FIG. 1, where such NMR plots are determined in accordance with another aspect of the present invention.

FIG. 3 is system diagram including a perceptual coder and decoder employing an embodiment of the present invention.

Detailed Description

Present inventive processing of input signals advantageously comprises three main functions: (i) determining the envelope of the part of the audio signal x(t) which is inside a particular cochlea filter band (or so called critical band), (ii) quantifying a roughness measure for the envelope, and (iii) mapping the roughness measure to a NMR for the part of the input signal. This process can then be repeated for determining NMRs of the sginal for each critical band. The analysis and methodology for each of these processing functions will now be explored in turn.

Signal envelope for a particular cochlea filter band

It has been shown, *e.g.*, in J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coding by using temporal noise shaping (TNS)," in *101st AES Convention, Los Angeles*, November 1996, that given a real, time domain signal, $x(t)$, the square of its Hilbert envelope, $e(t)$, can be expressed as

$$e(t) = F^{-1}\left\{\int \tilde{X}(\varepsilon)\cdot \tilde{X}^{*}(\varepsilon - f)\,d\varepsilon\right\} \tag{1}$$

If $X(f)$ is the Fourier transform of $x(t)$, then $\tilde{X}(f)$ is the Fourier transform of its analytic signal, and is a single sided frequency spectrum defined as

$$\tilde{X}(f) = \begin{cases} 0 & f < 0 \\ X(f) & f = 0 \\ 2X(f) & f > 0 \end{cases} \tag{2}$$

The signal envelope, which corresponds to the part of the signal that is inside a specific cochlea filter band, can be calculated by first filtering $\tilde{X}(f)$ of (1) by the cochlea filter, $H_i(f)$, *i.e.*,

$$\tilde{X}_i(f) = \tilde{X}(f) H_i(f). \tag{3}$$

Cochlea bands and filtering are described, *e.g.*, in J. B. Allen, "Cochlear micromechanics: A physical model of transduction," *JASA,* vol. 68, no. 6, pp. 1660-1670, 1980; and in J. B. Allen, "Modeling the noise damaged cochlea," in *The Mechanics and Biophysics of Hearing* (P. Dallos, C. D. Geisler, J. W. Matthews, M. A. Ruggero, and C. R. Steele, eds.), (New York), pp. 324-332, Springer-Verlag, 1991.

Thus, Eq. (1) can be re-written as:

$$e_i(t) = F^{-1}\left\{ \int \tilde{X}_i(\varepsilon) \cdot \tilde{X}_i^*(\varepsilon - f)\, d\varepsilon \right\} \tag{4}$$

In Eq. (4) $e_i(t)$ is the square of the signal envelope corresponding to the *i*th cochlea filter band whose characteristic frequency is f_i. F^{-1} in Eq. 4 represents the well-known Inverse Fourier Transform.

Quantifying envelope roughness

Eq. (1), or Eq. (4), shows that an input audio signal envelope may be derived from the autocorrelation function of its single sided frequency spectrum, $\tilde{X}(f)$. This relationship will be seen to be the dual of the following well-known formula which relates the power spectrum density of a signal, $S_{xx}(f)$, to is autocorrelation function in time domain:

$$S_{xx}(f) = F\left\{ \int x(\tau) \cdot x^*(\tau - t)\, d\tau \right\}, \tag{5}$$

where *F* denotes Fourier Transform.

By exploiting this duality, many well-established theories in time domain Linear Prediction (LP) processing can be applied to frequency domain. In particular, one well-known relationship between prediction gain and spectral flatness measure, described, for example, in N. S. Jayant and P. Noll, *Digital Coding of Waveforms – Principles and Applications to Speech and Video, page 56.* Prentice Hall, 1984, may be used to advantage. In accordance with such teachings, the rougher the frequency-domain spectrum $S_{xx}(f)$, the more predictable is the corresponding time signal $x(t)$; *i.e.*, the higher the prediction gain. (As is well known, prediction gain is defined as the ratio of original signal power to the power of the prediction residual error.)

Based on the duality of Eqs. (1) and (5), the following conclusion can be made: If linear prediction is applied to coefficients of $\tilde{X}(f)$, the single sided spectrum of the time signal $x(t)$, then a higher prediction gain corresponds to a rougher signal envelope $e(t)$. Therefore, for Eq. (4), prediction of $\tilde{X}_i(f)$ in the frequency domain serves as a reliable measure of the roughness of the signal envelope, $e_i(t)$. For an input signal comprising only white noise, prediction gain of its $\tilde{X}_i(f)$ will be the highest among all the signals, since it has the roughest envelope in time domain. On the other hand, prediction gain of $\tilde{X}_i(f)$ for pure tones will be the smallest, since they have flat a time domain envelope.

Linear Prediction (LP) operations are well-known and are described, for example in the above-cited book by Jayant and Noll at page 267. In the context of the present description, the input to LP operations is advantageously chosen as $\tilde{X}(f)$, rather than time-domain inputs, as is often the case.

Roughness of illustrative white noise and pure tone are shown in FIG. 1 on the traditional Bark scale. It should be noted that since the time signal is illustratively windowed by the well-known sin function (thereby increasing the roughness of the flat envelope of a pure tone), roughness of the illustrative pure tone is therefore greater than unity.

Calculate NMR from roughness

In accordance with an illustrative embodiment of the present invention, mapping a calculated roughness measure for an arbitrary signal to the NMR of the signal is advantageously accomplished using the following steps:

1. The calculated roughness measure of an arbitrary signal is normalized by that of a pure tone, since a pure tone has the flatest envelope.

2. Square the normalized roughness, since NMR is required in the signal energy domain.

3. The value obtained in step 2 is raised to the 4^{th} power to take into account the effect of the cochlea compression.

The resulting value is then directly proportional to the NMR of the signal. In other words, the signal NMR is calculated as follows:

$$NMR_i = c \cdot \left[\frac{r_s(i)}{r_t(i)} \right]^8, \qquad (6)$$

where r_s and r_t are the roughness of an arbitrary signal and a pure tone, respectively. Subscript, i denotes values for the ith cochlea filter band. In accordance with another aspect of the illustrative embodiment, the constant, c, is calculated by averaging its values for all i obtained by substituting $r_n(i)$ (the calculated roughness for a white noise input signal) for $r_s(i)$ and the theoretical NMR values.

The plot of NMRs for white noise shown in FIG. 2 support the accuracy of Eq. (6). That is, it is clear that the resulting NMRs are very close to their theoretical value of -6 dB, as discussed, *e.g.*, in R. P. Hellman, "Asymmetry in masking between noise and tone," *Perception and Psychophyics.*, vol. 11, pp. 241-246, 1972.

Illustrative System Overview

FIG. 3 shows a system organization for an illustrative embodiment of the present invention. In FIG. 3, an analog signal on input 300 is applied to preprocessor 305 where it is sampled (typically at 44.1 kHz) and each sample is converted to a digital sequence (typically 16 bits) in standard fashion. Of course, if input audio signals are presented in digital form, no such sampling and conversion is required.

Preprocessor 305 then advantageously groups these digital values in frames (or blocks or sets) of, *e.g.*, 2048 digital values, corresponding to, an illustrative 46 msec of audio input. Other typical values for these and other system or process parameters are discussed in the literature and known in well-known audio processing applications. Also, as is well known in practice, it proves advantageous to overlap contiguous frames, typically to the extent of 50 percent. That is, though each frame contains 2048 ordered digital values, 1024 of these values are repeated from the preceding 2048-value frame. Thus each input digital value appears in two successive frames, first as part of the second half of the frame and then as part of the first half of the frame. Other particular overlapping parameters are well-known in the art. These time-domain signal frames are then transformed in filterbank block 310 using. *e.g.*, a modified discrete cosine transform (MDCT) such as that described in J. Princen, *et al.*, "Sub-band Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," IEEE ICASSP,

1987, pp. 2161-2164. The illustrative resulting set of 1024 real coefficients (zero-frequency, Nyquist frequency, and all intermediate frequencies) resulting from the illustrative MDCT represents the short-term frequency spectrum of the input signal.

These MDCT coefficients are then quantized based on the NMRs calculated, illustratively using the method described above. Thus, by way of illustration:

1. For each frame (2048 samples resulted from block 305), calculate the Fourier Transform of its analytic signal, $\tilde{X}(f)$ defined in Eq. 2.
2. For the ith scale factor band (SFB), calculate $\tilde{X}_i(f)$ using Eq. 3, where the cochlear filter's ($H_i(f)$) characteristic frequency f_i is the center frequency of this particular scale factor band.
3. Perform Linear Prediction on $\tilde{X}_i(f)$ and denote its prediction gain as $r_s(i)$.
4. Use Eq. 6 to map the roughness of the signal in this SFB, $r_s(i)$, to NMR_i
5. Calculate the average signal power per frequency bin in this SFB, and then multiply it with NMR_i to get the scale factor for this SFB.
6. Quantize all MDCT coefficients in this SFB using the resulting scale factor.
7. Repeat step 2-6 for all SFBs.

Steps 1-5 illustratively correspond to the perceptual model block 310. Outputs of this block are scale factors for performing quantization in block 315 (step 6 above). All these scale factors will be sent as side information along with the quantized MDCT coefficients to medium 320.

Perceptual model block 310 shown in FIG. 3 includes the perceptual modeling improvements of the present invention described above in illustrative embodiments. Filter bank 308 is shown supplying frequency components for the respective SFB, i, to the quantizer/coder 315 and to perceptual model 310 for calculating the average signal power in the SFB (step 5). The NMR has to be calculated (step 1-5) from the corresponding time signal frame resulted from block 305.

Quantizer/coder block 315 in FIG. 3 represents well-known quantizer-coder structures that respond to perceptual model inputs and frequency components received from a source of frequency domain information, such as filter bank 308, for an input

signal. Quantizer/coder 315 will correspond in various embodiments of the present invention to the well-known AAC coder, but other applications of the present invention may employ various transform or OCF coders and other standards-based coders.

Block 320 in FIG. 3 represents a recording or transmission medium to which the coded outputs of quantizer/coder 315 are applied. Suitable formatting and modulation of the output signals from quantizer/coder 315 should be understood to be included in the medium block 320. Such techniques are well known to the art and will be dictated by the particular medium, transmission or recording rates and other system parameters. Further, if the medium 320 includes noise or other corrupting influences, it may be necessary to include additional error-control devices or processes, as is well known in the art. Thus, for example, if the medium is an optical recording medium similar to the standard CD devices, then redundancy coding of the type common in that medium can be used with the present invention. If the medium is one used for transmission, e.g., a broadcast, telephone, or satellite medium, then other appropriate error control mechanisms will advantageously be applied. Any modulation, redundancy or other coding to accommodate (or combat the effects of) the medium will, of course, be reversed (or otherwise subject to any appropriate complementary processing) upon the delivery from the channel or other medium 320 to a decoder, such as 330 in FIG. 3.

Coding parameters, including scale factors information used at quantizer/coder 315 are therefore sent as *side information* along with quantized frequency coefficients. Such side information is used in decoder 330 and perceptual decoder 340 to reconstruct the original input signal from input 300 and supply this reconstructed signal on output port 360 after performing suitable conversion to time-domain signals, digital-to-analog conversion and any other desired post-processing in unit 350 in FIG. 3. NMR side information is, of course supplied to perceptual decoder 340 for use there in controlling decoder 330 in restoring uniform quantization of transform (frequency) domain signals suitable for transformation back to the time domain.

The originally coded information provided by quantizer/coder 315 will therefore be applied at a reproduction device, *e.g.*, a CD player. Output on 360 is in such form as to be perceived by a listener upon playback as substantially identical to that supplied on input 100.

Those skilled in the art will recognize that numerous alternative embodiments of the present invention, and methods of practicing the present invention, in light of the present description.

What is claimed is: